A NOVEL METHOD FOR DIAGNOSIS OF BREAST CANCER USING K-MEANS ALGORITHM

SARANYA. V[1], SUBALAKSHMI. G [2], Dr HEMAJOTHI[3]
Final Year Students[1][2], Assistant Professor[3],
Department of Electronics & Communication Engineering,
Prince Shri Venkateshwara Padmavathy Engineering College, Chennai 127

Abstract

Breast cancer is an uncontrolled growth of breast cells. After skin cancer, breast cancer is the most commonly diagnosed cancer in women. Detection of mitotic cells is one of the critical markers for diagnosis of breast cancer. Detection and classification of breast cancer with high accuracy uses a novel Computer-Aided Detection (CAD) system for automatic diagnosing of benign/malignant breast tissues. A high complexity and low precision hybrid algorithm combining K-means and morphological operations is proposed in this study. The classical CAD for breast cancer helps in finding the Region of Interest (ROI) in mammogram images then, applying morphological operations, such as shape, texture, and density, to manually generate feature vectors and finally diagnosing benign and malignant tumors by classifying these feature vectors using KNN classifier. The combination of these measures that has higher advantages and a potential of highly accurate clustering accuracy.

Keywords: Breast cancer, Mammogram images, K-means clustering algorithm, Computer-Aided Detection (CAD), Morphological operations, KNN classifier.

1. Introduction

Canceris an abnormal growth of cells which tend to proliferate in an uncontrolled way.Cancer occurs when changes called mutations take place in genes that regulate cell growth. Breast cancer is cancer that forms inthe cells of the breasts.Typically, the cancer forms in either the lobules or the ducts of the breast. In generally, pathology experts manually mark the mitotic cell on High power fields (HPF). However, manual annotation of mitosis is very time-consuming task since a single whole slide image may contain a large number of HPF. Besides due to the high biological variation of mitotic cell, manual detection is usually prone to error. Thus, Mammography screening is the only method presently considered appropriate for mass screening of asymptomatic women. This paper provides a conception and implementation of Computer Assisted Detection (CAD) for mammogram images classification.

The machine learning algorithms is used for breast cancer classification. The right analysis of Breast Cancer and categorization of patients into malignant and non-malignant group is the subject of much examination. Machine learning techniques are convincing methods to characterize information [11].Abdulsalam Alarabeyyat, Mohannad Alhanahnah etal.[10] haveproposed a new method to detect the breast cancer with high accuracy. This method consists of two main parts, in the first part the image processing techniques are used to prepare the mammography images for feature and pattern extraction process. The second part is presented by utilizing the extracted features as an input for supervised learning models.

Image processing is the use of a digital computer to process digital images through an algorithm. It is a method that develops to convert the image into digital form and perform some operations to obtain specific models or to extract useful information from it. By the use of image processing techniques, it has become easy to detect cancerous mass from an infected breast. The aim of pre-processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing. Data sets can require pre-processing techniques to ensure accurate, efficient, or meaningful analysis. Non-linear cascading filters are used to remove salt and pepper noise.

K-means clustering is used to compare the result based on test data. As a result, a set of genes are identified that are potential bio marks for breast cancer prognosis which can categorize the patients based on the certain attributes [13].Segmentation is a process of grouping together pixels that have similar attributes. The image goes through thresholding process for the purpose of segmenting the ROI of the image. K- means algorithm is an iterative algorithm that partitions the dataset according to their features into K number of predefined non- overlapping distinct clusters or subgroups. KNN is one of the classification algorithms which uses the entire dataset in its training phase.

2. Related works

Convolutional neural network (CNN) is used for automatically extract mitosis features. The region proposed network (RPN) to locate a set of class-agnostic mitosis proposals. The improved R-CNN subnet to screen for mitosis from these proposals [6]. Utilizing the extracted features as an input for a two types of supervised learning models, which are Back Propagation Neural Network (BPNN) model and the Logistic Regression (LR) model [10]. K-means clustering is used to compare the result based on test data. As a result, a set of genes are identified that are potential bio marks for breast cancer prognosis which can categorize the patients based on the certain attributes [13].

An automatic method for detecting mitosis. The mitosis detection task as a semantic segmentation problem and use a deep fully convolutional network to address it [3]. A deep learning algorithm that can accurately detect breast cancer on screening mammograms using an "end-to-end" training approach that efficiently leverages training datasets [9]. A hybrid approach based on mad normalization, KMC based feature weighting and AdaBoostM1 classifier. the AdaBoostM1 classifier has been used to classify the weighted data set [7]. An approach that improves the accuracy and enhances the performance of three different classifiers: Decision Tree, Naive Bayes (NB), and Sequential Minimal Optimization (SMO) [15].

A novel deep learning framework for the detection and classification of breast cancer in breast cytology images using the concept of transfer learning [14]. The two most popularly used Supervised Machine Learning Algorithms, K-Nearest Neighbour and Naive Bayes has achieved a best accuracy of 97.15% by employing the KNN algorithm and a lowest error rate of 96.19% using NB classifier [2]. The right analysis of Breast Cancer and categorization of patients into malignant and non-malignant group is the subject of much examination. Machine learning techniques are thus the convincing methods [11].

A genetically optimized neural network (GONN) for breast cancer classification (malignant and benign). They optimized the neural network architecture by introducing new crossover and mutation operators [1]. A conception and implementation of Computer Assisted Detection (CAD) for mammogram images classification. The system is based on a GA-based features selection algorithm to reduce the dimensionality of the feature vector [12]. An SVM-based ensemble learning model for breast cancer diagnosis. The proposed ensemble model includes two types of SVM structures, i.e., a C-SVM and a SVM, and six types of kernel functions [5].

A method based on the extraction of image patches for training the Convolutional Neural Network (CNN) and the combination of these patches for final classification [4]. A robust machine learning classification techniques such as Support vector machine (SVM) kernels and Decision Tree to distinguish cancer mammograms from normal subjects [8].

3. Proposedmethod for Breast cancer detection

A novel Computer-Aided Detection (CAD) system is used to reduce the human factor involvement and to help the radiologist in automatic diagnosis of benign/malignant breast tissues by utilizing the Basic morphological operations. The input Region of Interest (ROI) is extracted manually and subjected to further number of pre-processing stages. ROI in an input frame is known as ROI Segmentation. In ROI Segmentation, it selects a specific region in the frame. The geometrical and texture features are extracted for feature extraction of suspicious region.

A. Input

It reads and displays an input Image. In image processing, input is defined as the action of retrieving an image from some source, usually a hardware-based source for processing. It is the first step in the workflow sequence because, without an image, no processing is possible.

B. Pre-processing

Data sets can require preprocessing techniques to ensure accurate, efficient, or meaningful analysis. This technique consists of resize the input image and converting the input image into gray scale image and using filters. Data cleaning refers to methods for finding, removing, and replacing bad or missing data. Smoothing and detraining are processes for removing noise and linear trends from data, while scaling changes the bounds of the data.

C. Segmentation

The technique of partitioning the image into segment can be defined as image segmentation. Considering the similar property, segmentation is implemented. This similar property is cluster together approach that implements the k-mean clustering algorithm by introducing repeated segmentation scheme which explores the centroid of each set in the segment and eventually re-segment the input based on the closest centroid. This technique aids in the extraction of important image characteristics, based on which information can be easily perceived. Morphological operations like DILATION, EROSION, AREA OPENING, CLOSING, BORDER CLEARING is used for segmentation process.

K- means algorithm is an iterative algorithm that partitions the dataset according to their features into K number of predefined non- overlapping distinct clusters or subgroups. It allocates the data points to a cluster if the sum of the squared distance between the cluster's centroid and the data points is at a minimum where the cluster's centroid is the arithmetic mean of the data points that are in the cluster.

D. **Feature Extraction**

Transforming the input data into the set of features is called feature extraction. Feature extraction involves simplifying the number of resources required to describe a large set of data accurately. Texture based feature extraction method like GLCM (Grey level co-occurrence matrix) is used for feature extraction. The glum gives the texture features of the test image like contrast, correlation, energy and etc. Then the region-based features give the various different features of the input image like area, diameter etc. The best features that are related is used to differentiate the Benign and malignant cancers.

E. KNN Classification

The KNN (K-Nearest Neighbor)gives more accurate data classification which is used to select k as an odd number which avoids the irregular data. The KNN procedure is the technique used in ML procedures. Classically, Euclidean distance is used as the distance metric. However, this is only suitable for endless variables. KNN is a new process that deliveries all available cases and categorizes novel cases built on an evaluation quantity (e.g., distance functions). KNN procedure is identical simple. It works built on a minimum distance from the interrogation instance to the training samples to regulate the K-nearest neighbors. The information for KNN procedure contains numerous attribute which will be used to categorize. The information of KNN can be any dimension scale from insignificant, to measurable scale. Figure 3.1 shows theblock diagram of detection of breast cancer using K-Means algorithm.

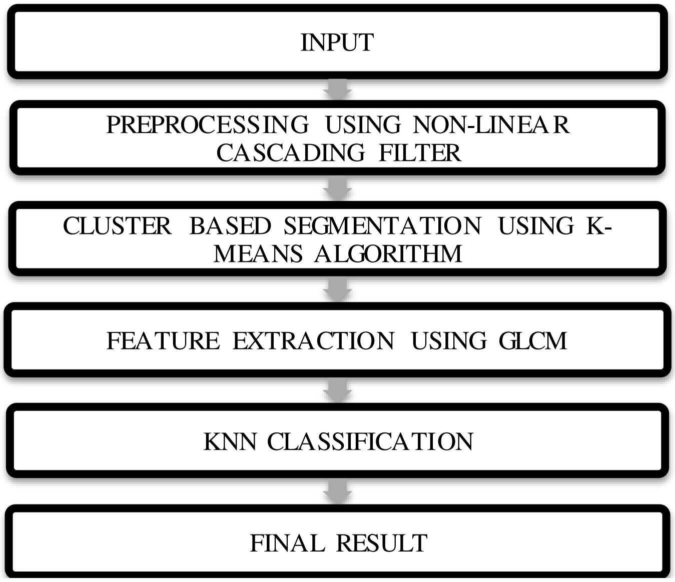
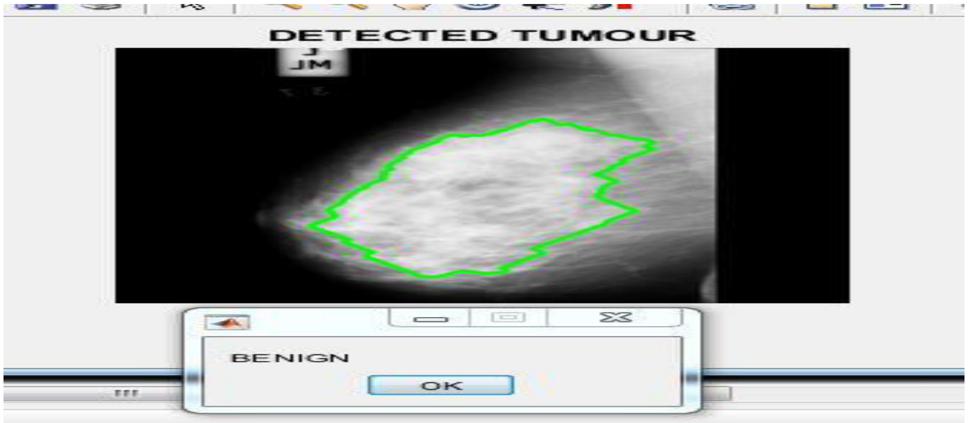

INPUT
PREPROCESSING USING NON-LINEAR CASCADING FILTER
CLUSTER BASED SEGMENTATION USING K-MEANS ALGORITHM
FEATURE EXTRACTION USING GLCM
KNN CLASSIFICATION
FINAL RESULT

Figure 3.1 Block diagram of detection of breast cancer

4. Resultsand discussion

The results were obtained by using k-means algorithm and are discussed with different cases. The input image is obtained and is pre-proposed to remove the noises. Each image in the dataset has been resized and have resolution of 1024x1024. MATLAB R2018a is used to perform the proposed idea and it is tested using several images. The implementation reveals the success of the system in classifying the database by up to 97.461% and also classify the cancerous stage using KNN classifier.



5. Conclusion

An optimized KNN model is proposed for breast cancer prediction. The algorithm K-nearest neighbors is used for breast cancer classification. The algorithm with the parameter K is used as a variable factor. This algorithm has classified the cancer with 97.461% accuracy. Compared to the previous existing method, which has manual detection method,the proposed system reduces the loss of life. In future, the accuracy is enhanced.

References

1. Arpit Bhardwaj, Aruna Tiwari, "Breast Cancer Diagnosis Using Genetically Optimized Neural Network Model", 2015.
2. Bakthavachalam.B.D, Dr Albert Antony Raj.S., "A Study of Breast Cancer Analysis Using K-Nearest Neighbor with Different Distance Measures and Classification Rules Using Machine Learning", 2020.
3. Chao Li, Xingang Wang, Wenyu Liu, Longin Jan Latecki, Bo Wang c, Junzhou Huang,"Weakly supervised mitosis detection in breast histopathology images using concentric loss", 2019.
4. Fabio Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte, "Breast Cancer Histopathological Image Classification using Convolutional Neural Networks", International Joint Conference on Neural Networks (IJCNN), 2016.
5. Haifeng Wang, Bichen Zheng, Sang Won Yoon,Hoo Sang Ko, "A Support Vector Machine-Based Ensemble Algorithm for Breast Cancer Diagnosis", 2016.
6. Hai Jun Lei1, Shamoun Liu1, Hai Xie2, Jong Yih Kuo3, and Baiying Lei4, "An Improved Object Detection Method for Mitosis Detection", 2019.
7. Kemal Polat , Umit Senturk , "A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier", 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), IEEE, 2018.
8. Lal Hussain, Wajid Aziz, Sharjil Saeed, Saima Rathore, Muhammad Rafique, "Automated Breast Cancer Detection Using Machine Learning Techniques by Extracting Different Feature Extracting Strategies", 2018.
9. Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride and Weiva Sieh, "Deep Learning to Improve Breast Cancer Detection on Screening Mammography", 2019.
10. Moh'd Hadidi, Abdulsalam Alarabeyyat, and Mohannad Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm", 2016.
11. Nandita Goyal, Munesh Chandra Trivedi, "Breast cancer classification and identification using machine learning approaches", 2019.
12. Nawel Zemmal, Nabiha Azizi, Nilanjan Dey, Mokhtar Sellami, "Adaptive Semi Supervised Support Vector Machine Semi Supervised Learning with Features Cooperation for Breast Cancer Classification", 2016.
13. Radha. R, Rajendiran .P, "Using K-Means Clustering Technique to Study of Breast Cancer", 2019.
14. Sana Ullah Khan, Naveed Islam, Ikram Ud Din, Zahoor Jan, Joel J. P. C Rodrigues , "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning", 2019.

15. Siham A. Mohammed, Sadeq Darrab, Siham A. MohammedSadeq Darrab, “Analysis of Breast Cancer Detection Using Different Machine Learning Techniques”, 2020.